Title:

Subtitle

May 22, 2024

Abstract

1 Introduction

The National Institute of Standards and Technology Artificial Intelligence (AI) Risk Management Framework (RMF).[**?**]

2 Generative AI Governance

3 Generative AI Inventories

4 Generative AI Risk Tiers

5 Generative AI Risk Measurement

6 Generative AI Risk Management

Conclusion

Acknowledgments

Thank you to Bernie Siskin and Nick Schmidt of BLDS and Eric Sublett of Relman Colfax for formative discussions relating to GAI risk tiering.

Abbreviations

- AI: Artificial Intelligence
- AI RMF: Artificial Intelligence Risk Management Framework
- GAI: Generative AI
- RMF: Risk Management Framework

Appendix A: Example Generative AI–Trustworthy Characteristic Crosswalk

6.1 A.1: Trustworthy Characteristic to Generative AI Risk Crosswalk

Table 1: Trustworthy Characteristic to Generative AI Risk Crosswalk.

Accountable and Transparent	Explainable and Interpretable	Fair with Harmful Bias Managed	Privacy Enhanced
Data Privacy	Human-AI Configuration	Confabulation	Data Privacy
Environmental	Value Chain and Component Integration	Environmental	Human-AI Configuration
Human-AI Configuration		Human-AI Configuration	Information Security
Information Integrity		Intellectual Property	Intellectual Property
Intellectual Property		Obscene, Degrading, and/or Abusive Content	Value Chain and Component Integration
Value Chain and Component Integration		Toxicity, Bias, and Homogenization	
		Value Chain and Component Integration	

Safe	Secure and Resilient	Valid and Reliable
CBRN Information	Dangerous or Violent Recommendations	Confabulation
Confabulation	Data Privacy	Human-AI Configuration
Dangerous or Violent Recommendations	Human-AI Configuration	Information Integrity
Data Privacy	Information Security	Information Security
Environmental	Value Chain and Component Integration	Toxicity, Bias, and Homogenization
Human-AI Configuration		Value Chain and Component Integration
Information Integrity		
Information Security		
Obscene, Degrading, and/or Abusive Content		
Value Chain and Component Integration		

6.2 A.2: Generative AI Risk to Trustworthy Characteristic Crosswalk

Table 2: Generative AI Risk to Trustworthy Characteristic Crosswalk.

CBRN Information	Confabulation	Dangerous or Violent Recommendations	Data Privacy
Safe	Fair with Harmful Bias Managed Safe Valid and Reliable	Safe Secure and Resilient	Accountable and Transparent Privacy Enhanced Safe Secure and Resilient

Environmental	Human-AI Configuration	Information Integrity	Information Security
Accountable and Transparent Fair with Harmful Bias Managed Safe	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable	Accountable and Transparent Safe Valid and Reliable	Privacy Enhanced Safe Secure and Resilient Valid and Reliable

Intellectual Property	Obscene, Degrading, and/or Abusive Content	Toxicity, Bias, and Homogenization	Value Chain and Component Integration
Accountable and Transparent Fair with Harmful Bias Managed Privacy Enhanced	Fair with Harmful Bias Managed Safe	Fair with Harmful Bias Managed Valid and Reliable	Accountable and Transparent Explainable and Interpretable Fair with Harmful Bias Managed Privacy Enhanced Safe Secure and Resilient Valid and Reliable

Appendix B: Example Risk Tiers for Generative AI

Appendix C: List of Publicly Available Model Testing Suites ("Evals")

C.1: Publicly Available Model Testing Suites ("Evals") by Trustworthy Characteristic

Table 3: Publicly Available Model Testing Suites ("Evals") by Trustworthy Characteristic.

Accountable and Transparent
An Evaluation on Large Language Model Outputs: Discourse and Memorization (see Appendix B)
Big-bench: Truthfulness
DecodingTrust: Machine Ethics
Evaluation Harness: ETHICS
HELM: Copyright
Mark My Words

Fair with Harmful Bias Managed
BELEBELE
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
C-Eval (Chinese evaluation suite)
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
Finding New Biases in Language Models with a Holistic Descriptor Dataset
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models
HELM: Bias
HELM: Toxicity
MT-bench
The Self-Perception and Political Biases of ChatGPT
Towards Measuring the Representation of Subjective Global Opinions in Language Models

Privacy Enhanced
HELM: Copyright
llmprivacy
mimir

Safe
Big-bench: Convince Me
Big-bench: Truthfulness
HELM: Reiteration, Wedging
Mark My Words
MLCommons
The WMDP Benchmark

Publicly Available Model Testing Suites ("Evals") by Trustworthy Characteristic (continued).

Secure and Resilient
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations
detect-pretrain-code
In-The-Wild Jailbreak Prompts on LLMs
JailbreakingLLMs
llmprivacy
mimir
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs

Valid and Reliable
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Context Free Question Answering
Big-bench: Contextual question answering, Reading comprehension, Question generation
Big-bench: Morphology, Grammar, Syntax
Big-bench: Out-of-Distribution
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Reading comprehension
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: BLiMP
Evaluation Harness: CoQA, ARC
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness
FLASK: Readability, Conciseness, Insightfulness
HELM: Knowledge
HELM: Language
HELM: Text classification
HELM: Question answering
HELM: Reasoning
HELM: Robustness to contrast sets
HELM: Summarization
Hugging Face: Fill-mask, Text generation - Benchmarking
Hugging Face: Question answering
Hugging Face: Summarization
Hugging Face: Text classification, Token classification, Zero-shot classification
MASSIVE
MT-bench

C.2: Publicly Available Model Testing Suites ("Evals") by Generative AI Risk

Table 4: Publicly Available Model Testing Suites ("Evals") by Generative AI Risk.

CBRN Information
Big-bench: Convince Me
Big-bench: Truthfulness
HELM: Reiteration, Wedging
MLCommons
The WMDP Benchmark

Confabulation
BELEBELE
Big-bench: Algorithms, Logical reasoning, Implicit reasoning, Mathematics, Arithmetic, Algebra, Mathematical proof, Fallacy, Negation, Computer code, Probabilistic reasoning, Social reasoning, Analogical reasoning, Multi-step, Understanding the World
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Context Free Question Answering
Big-bench: Contextual question answering, Reading comprehension, Question generation
Big-bench: Convince Me
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Morphology, Grammar, Syntax
Big-bench: Out-of-Distribution
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
Big-bench: Truthfulness
C-Eval (Chinese evaluation suite)
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet Reading comprehension
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: BLiMP
Evaluation Harness: CoQA, ARC
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness
FLASK: Readability, Conciseness, Insightfulness
Finding New Biases in Language Models with a Holistic Descriptor Dataset
HELM: Knowledge
HELM: Language
HELM: Language (Twitter AAE)
HELM: Question answering
HELM: Reasoning
HELM: Reiteration, Wedging
HELM: Robustness to contrast sets
HELM: Summarization
HELM: Text classification
Hugging Face: Fill-mask, Text generation
Hugging Face: Question answering
Hugging Face: Summarization
Hugging Face: Text classification, Token classification, Zero-shot classification
MASSIVE
MLCommons
MT-bench

Publicly Available Model Testing Suites ("Evals") by Generative AI Risk (continued).

Dangerous or Violent Recommendations
Big-bench: Convince Me
Big-bench: Toxicity
DecodingTrust: Adversarial Robustness, Robustness Against Adversarial Demonstrations
DecodingTrust: Machine Ethics
DecodingTrust: Toxicity
Evaluation Harness: ToxiGen
HELM: Reiteration, Wedging
HELM: Toxicity
MLCommons

Data Privacy
An Evaluation on Large Language Model Outputs: Discourse and Memorization (with human scoring, see Appendix B)
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
DecodingTrust: Machine Ethics
Evaluation Harness: ETHICS
HELM: Copyright
In-The-Wild Jailbreak Prompts on LLMs
JailbreakingLLMs
MLCommons
Mark My Words
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs
detect-pretrain-code
llmprivacy
mimir

Information Integrity
Big-bench: Analytic entailment, Formal fallacies and syllogisms with negation, Entailed polarity
Big-bench: Convince Me
Big-bench: Paraphrase
Big-bench: Sufficient information
Big-bench: Summarization
Big-bench: Truthfulness
DecodingTrust: Machine Ethics
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Language Understanding
Eval Gauntlet: World Knowledge
Evaluation Harness: CoQA, ARC
Evaluation Harness: ETHICS
Evaluation Harness: GLUE
Evaluation Harness: HellaSwag, OpenBookQA, TruthfulQA
Evaluation Harness: MuTual
Evaluation Harness: PIQA, PROST, MC-TACO, MathQA, LogiQA, DROP
FLASK: Logical correctness, Logical robustness, Logical efficiency, Comprehension, Completeness
FLASK: Readability, Conciseness, Insightfulness
HELM: Knowledge
HELM: Language
HELM: Question answering
HELM: Reasoning
HELM: Reiteration, Wedging
HELM: Robustness to contrast sets
HELM: Summarization
HELM: Text classification
Hugging Face: Fill-mask, Text generation
Hugging Face: Question answering
Hugging Face: Summarization
MLCommons
MT-bench
Mark My Words

Publicly Available Model Testing Suites ("Evals") by Generative AI Risk (continued).

Information Security
Big-bench: Convince Me
Big-bench: Out-of-Distribution
Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation
DecodingTrust: Out-of-Distribution Robustness, Adversarial Robustness, Robustness Against Adversarial Demonstrations
Eval Gauntlet: Commonsense reasoning, Symbolic problem solving, Programming
HELM: Copyright
In-The-Wild Jailbreak Prompts on LLMs
JailbreakingLLMs
Mark My Words
TAP: A Query-Efficient Method for Jailbreaking Black-Box LLMs
detect-pretrain-code
llmprivacy
mimir

Intellectual Property
An Evaluation on Large Language Model Outputs: Discourse and Memorization (with human scoring, see Appendix B)
HELM: Copyright
Mark My Words
llmprivacy
mimir

Obscene, Degrading, and/or Abusive Content
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
HELM: Bias
HELM: Toxicity

Toxicity, Bias, and Homogenization
BELEBELE
Big-bench: Low-resource language, Non-English, Translation
Big-bench: Out-of-Distribution
Big-bench: Social bias, Racial bias, Gender bias, Religious bias
Big-bench: Toxicity
C-Eval (Chinese evaluation suite)
DecodingTrust: Fairness
DecodingTrust: Stereotype Bias
DecodingTrust: Toxicity
Eval Gauntlet: World Knowledge
Evaluation Harness: CrowS-Pairs
Evaluation Harness: ToxiGen
Finding New Biases in Language Models with a Holistic Descriptor Dataset
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models
HELM: Bias
HELM: Toxicity
The Self-Perception and Political Biases of ChatGPT
Towards Measuring the Representation of Subjective Global Opinions in Language Models